Safe Reinforcement Learning Using Advantage-Based Intervention

Nolan Wagener, Byron Boots, Ching-An Cheng

Georgia Tech | University of Washington | Microsoft Research

Safe RL Problem

Reward $r(s,a) \geq 0$ → Value $V^\pi(s)$
Cost $c(s,a) = \mathbf{1}\{s = \text{violation}\}$ → Value $\overline{V}^\pi(s)$

Goal: Maximize return while keeping cost below some threshold, *including during training*.

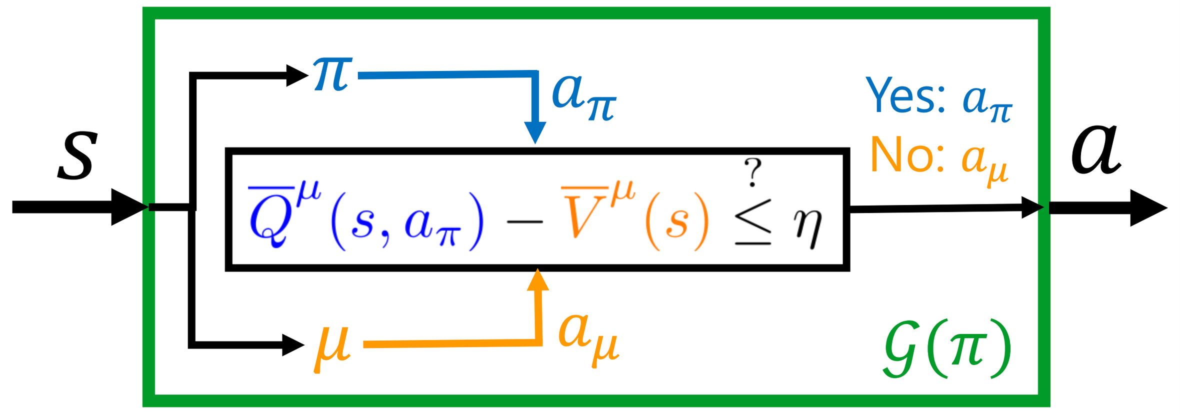
$$\max_\pi \; V^\pi(s_0) \quad \text{subject to} \quad \overline{V}^\pi(s_0) \leq \delta$$

Dilemma: Partially optimized RL policy π may be unsafe.
Assumption: Given baseline policy μ which is safe starting from the initial state s_0.
Solution: Use μ to prevent unsafe actions from π.
Dilemma: Optimized π needs to be safe and have high returns.
Solution: Augment original MDP with penalizing rewards for being intervened by μ.

Advantage-Based Intervention

Intervention rule $\mathcal{G}$ given by μ and threshold η.

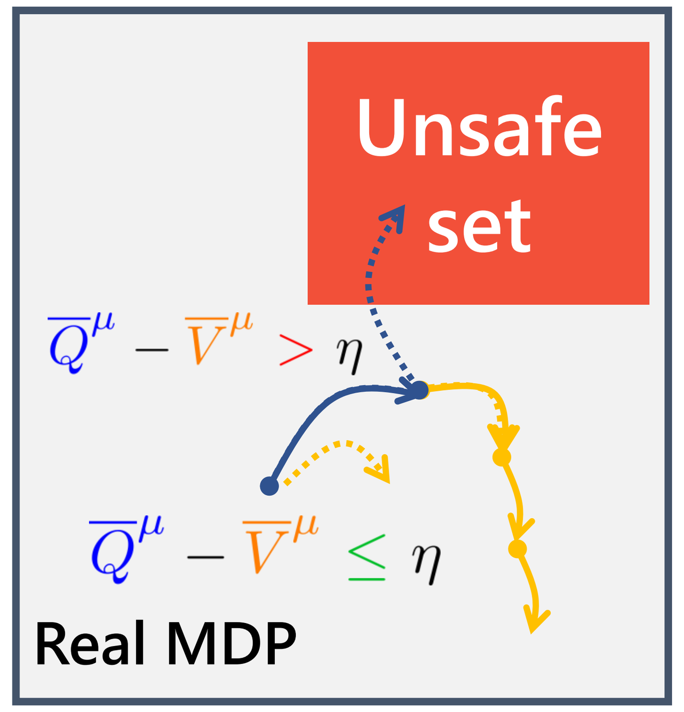
Shielded policy $\mathcal{G}(\pi)$ uses advantage $\bar{A}^\mu(s,a)$ w.r.t. μ to determine to sample from π or μ.

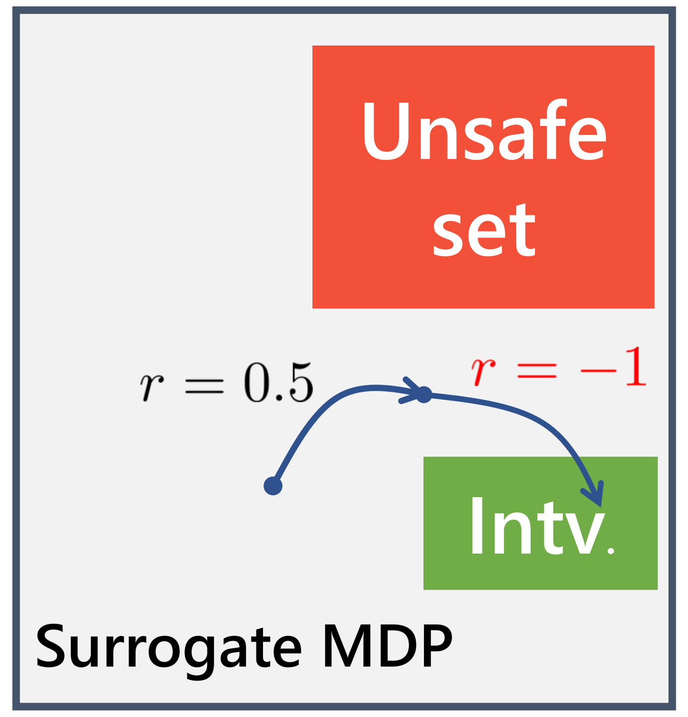


What if we don't have access to $\bar{A}^\mu(s,a)$?
Can learn approximation from data collected from μ.
Why intervene based on advantages instead of Q?
Allows reduction from constrained RL to *unconstrained* RL.

Safe RL with SAILR

Run **shielded policy** $\mathcal{G}(\pi)$ in real MDP.
From π's perspective, if intervened it transitions to absorbing state and gets a penalizing reward.
Run regular RL algorithm (e.g., PPO) on surrogate MDP.





Theoretical Results

Theorem (Safety During Training)
Shielded policy $\mathcal{G}(\pi)$ is nearly as safe as μ.

$$\overline{V}^{\mathcal{G}(\pi)}(s_0) \leq \overline{V}^\mu(s_0) + \frac{\eta}{1-\gamma}$$

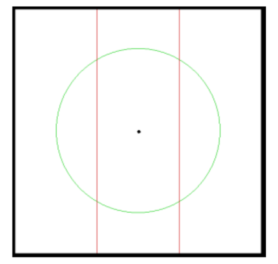
Theorem (Safety and Returns at Deployment)
SAILR policy $\hat{\pi}$ is nearly as safe as μ and has nearly the same returns as comparator policy π^*.

$$\overline{V}^{\hat{\pi}}(s_0) \leq \overline{V}^\mu(s_0) + \frac{\eta}{1-\gamma}$$

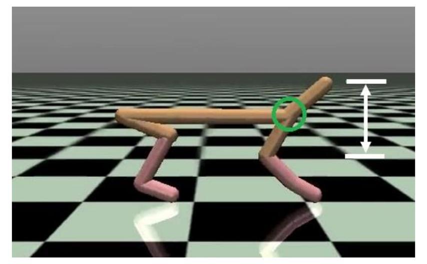
$$V^{\pi^*}(s_0) - V^{\hat{\pi}}(s_0) \leq O\left(\frac{\text{Prob}(\pi^* \text{ is intervened by } \mathcal{G})}{1-\gamma}\right)$$

Experiments

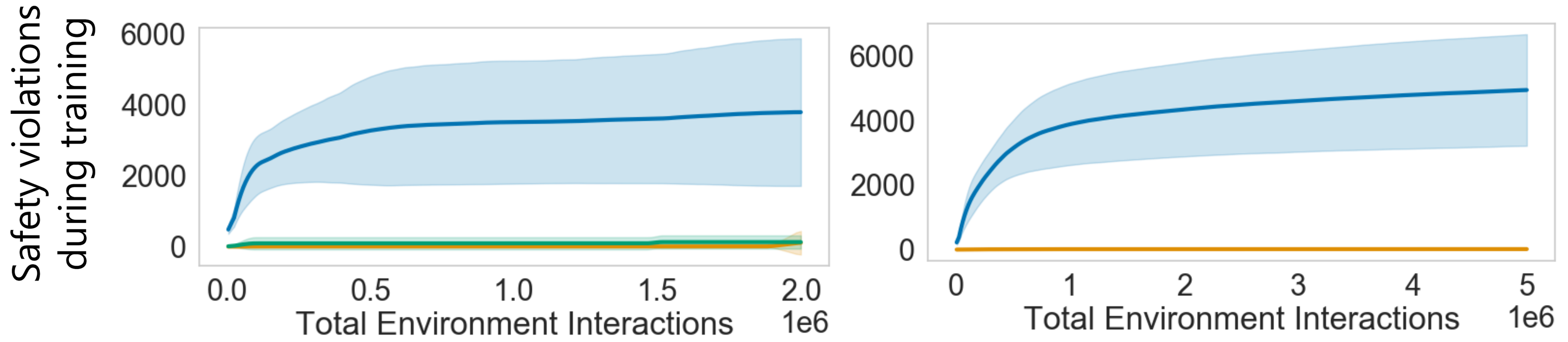
Point Robot
Reward: move fast in CCW
Constraint: don't touch vertical red lines
μ: deceleration policy



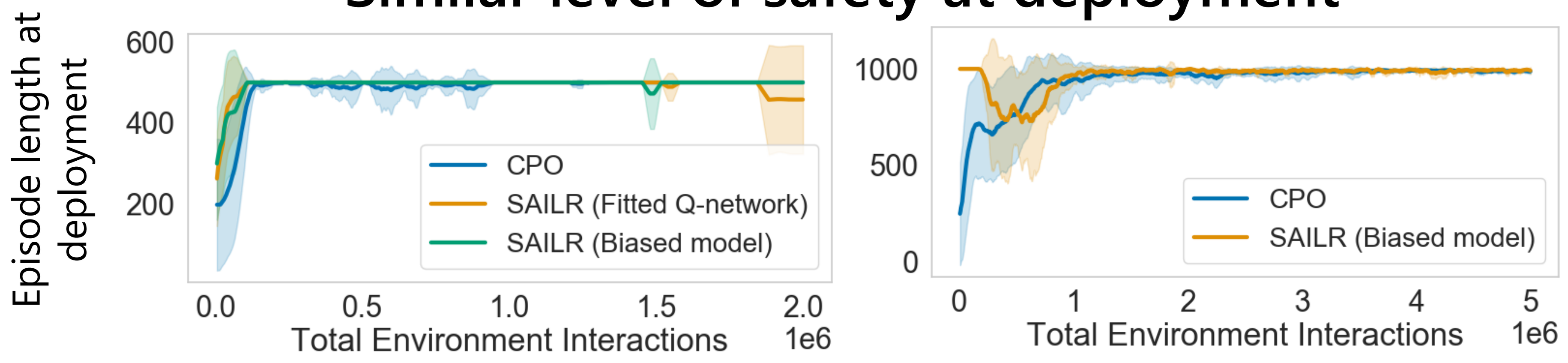
Half-Cheetah
Reward: run forward fast
Constraint: keep "chin" joint in a height range
μ: model predictive control



Far fewer safety violations during training



Similar level of safety at deployment



Similar returns at deployment